International Journal of Computer Science Engineering
and Information Technology Research(IJCSEITR)
ISSN(P): 2249-6831; ISSN(E): 2249-7943
Vol. 5, Issue 2, Apr 2015, 71-78

A GRAPH CLUSTERING APPROACH TO IDENTIFY KEY GENES FROM MICROARRAY GENE EXPRESSION DATA

JAHIRUDDIN

Department of Computer Science, Jamia Millia Islamia (Central University), New Delhi, India

ABSTRACT

To identify genes that play an important role in a disease is an important task. In this paper we proposed a method to identify key genes. For this first we analyze the genes based on their expression values using statistical technique and sort then in ascending order of their p-value. Then we construct gene regulatory network from microarray gene expression profile using correlation technique. In gene regulatory network construction, we take only those genes whose p-value is less than 0.01 and have the gene name. Thereafter we download the expression values of these genes in different samples. These expression values are further modified by applying Singular Value decomposition. The gene regulatory network is generated calculating interaction between genes using modified expression values. The interaction between filtered genes is calculated using correlation. Thereafter the graph is clustered by using Markov Cluster Algorithm and the gene corresponding to nodes of maximum degree for different value of parameter r is identified as key genes. Finally, validation of the result is done using existing literature.

KEYWORDS: Microarray Gene Expression, Gene Regulatory Network, Singular Value Decomposition, Latent Semantic Analysis, Markov Cluster Algorithm

1. INTRODUCTION

A gene is the basis for heredity. Human Genes made up of DNA (deoxyribonucleic acids) that stores the instruction to make the proteins. Human genes present in pair that comes from their both parents. Every person has same genes only very small percent of genes in persons are differs. Due to different of this small percentage of gene the property of each person is unique. Genes play important role in some fatal diseases. The identification of the gene responsible for such disease may play very import role in its treatment.

Recently microarray technique grows rapidly. By using this technique the analysis of expression of the thousands genes can be done in a single experiment with small sample size [11]. The microarray technology allows us to measure the expression levels of a large number of genes simultaneously in a microarray experiment [9]. A number of applications like disease diagnosis, drug discovery, gene discovery, and toxicogenomics are using microarray technique. Recently microarray technology has been extensively used by the researchers to identify key genes. There is a huge amount of microarray data, but it is scattered and is not available for public use. A large number of microarray data from various sources are stored in the Gene Expression Omnibus (GEO) data repository of the National Center for Biotechnology Information (NCBI). The Gene Expression Omnibus (GEO) also has the GEO2R web tool for comparing two or more group of samples to identify genes that are differently expressed. It performs comparisons on original processed data using the GEO query and limma R package. It performs a number of statistical testing based on corrected p-value.

A microarray gene expression data may be represented by a real matrix A of order m x n, where m is number of gene, n is number of samples, and a_{ij} is expression value of i^{th} gene in j^{th} sample. Microarray data have dimensionality (m >> n) and noise problems. Due to experimental limitation there is no control on these problems.

The prediction of key gene from microarray gene expression is a new technique. A gene regulatory network (GRN) is a collection of DNA segments in a cell is used to govern gene expression levels of mRNA and proteins. It is used to model regulatory interaction in the cell and represent the gene regulation. Microarray data can be used to construct gene regulatory network. The expression value of the gene in different samples may be used to calculate the interaction between gene pairs that may be used to construct the GRN. The mapping of the topology of gene regulatory networks is a central problem in system biology [8]. Therefore, accurate computational method needed for construction of GRN from gene expression profiles. One method of reconstructing regulatory networks from gene expression microarray data is information theoretic approach [12]. But correlation method of reconstruction of gene regulatory network is simple and fast.

Singular value decomposition (SVD) is a factorization technique that factorizes a rectangular real or complex matrix [4]. It is based on a linear algebra theorem that a rectangular matrix A of order m × n can be decomposed into three matrices – an orthogonal matrix U of order m × n, a diagonal matrix S of order n × n, and an orthogonal matrix V of order n × n such that $A=USV^*$, where V^* is the conjugate transpose of V. The columns of the matrix U are orthogonal eigenvectors of AA^T. Similarly, columns of V are orthogonal eigenvectors of A^TA, and S is a diagonal matrix containing the square roots of eigenvalues from U or V in descending order. SVD has the added benefit that the representation of genes that share samples become more similar to each other, and genes that were dissimilar may become more dissimilar. Presently, SVD is included in a number of software and java classes like MATLAB, JAMA java package etc. Caporaso et al. [2] explored Latent Semantic Analysis (LSA) that uses the SVD, for biomedical question answering system. They explained that LSA increases the number of phrases returned in response to a question.

There are a number of computational methods for GRN modelling in the literature. It includes Boolean networks, graph method, Bayesian networks, differential equations and so on [5], [6], [8], [10]. Madhamshettiwar et al. reconstructed cancer specific GRN for study [7]. They study application of gene regulatory network inference to ovarian cancer. An exhaustive state of art study for GRN modelling is done in [7]. They applied best method to infer GRN of ovarian cancer. In [1] information theoretic approach is used to construct gene regulatory network. They calculate the mutual information between genes to get the interaction between genes from gene expression profile. In [12] they proposed a fast method to calculate pair wise mutual information for gene regulatory network reconstruction.

In this paper, first we analyse microarray expression data using GEO2R online tool, then the genes and its expression values in samples are downloaded. Thereafter, correlation method is used to construct GRN. This network is further processed using Markov Cluster Algorithm (MCL) then the gene corresponding to hub nodes are marks as hub genes. The MCL is an iterative method that interleaves matrix expansion and inflation steps [13]. Matrix expansion corresponds to taking successive powers of the transition matrix, while matrix inflation makes the higher probability transition and reduces the lower probability transition. MCL does not require the number of clusters k in advance; it requires the inflation parameter r > 1. A small value of r results in small number of clusters of larger size, whereas a high value of r generates large number of clusters of smaller size. If r should be too large then every nodes should be isolated.

The remaining paper is organized as follows. Section 2 present the proposed system, experimental setup and results are described in section 3. Section 4 concludes the paper with future enhancement.

2. PROPOSED SYSTEM

We now present the complete architecture of our system, which is designed to identify key genes from microarray gene expression profile data. The key genes may be used in diagnostic and treatment of a disease. The proposed system is shown in figure 1. This is characterized by following key functionalities – *Microarray data loading and analysis, Genes and their expression values extraction, relationship identifying between gene pairs and GRN generation, GRN clustering* and *key gene identification* and *verification*. A brief description about these functionalities is given in the following paragraphs:

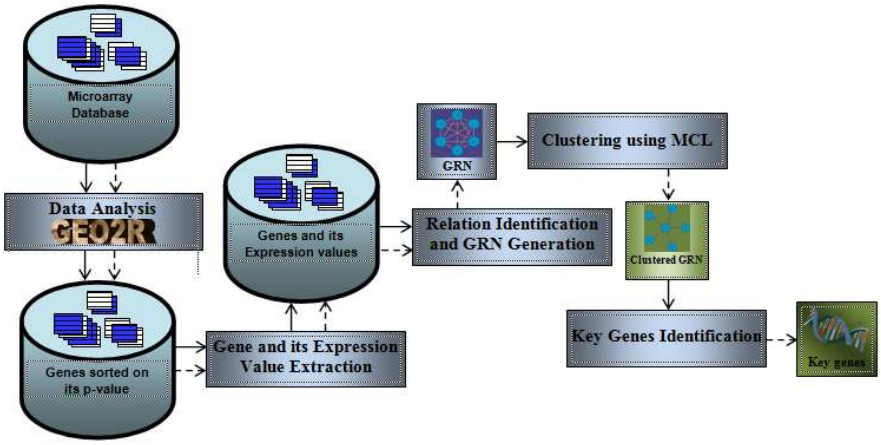

Figure 1: System Architecture

- **Microarray Data Loading and Analysis** is focused to download the microarray data and analyse them. The GEO's accession viewer interface may be used for this purpose. To download microarray data GEO accession number of that microarray data is needed. The GEO accession number of required microarray data is entered in the text field and go button is clicked to download the data. It returns the microarray data with its description. The analysis of the microarray data may be done by using GEO2R tool. This tool does t and F testing on the data and arranges the genes in ascending order of their p-values.
- **Genes and their Expression Values Extraction** is focused on extraction of the efficient genes and downloading their expression values. The extraction of efficient genes is based on their p-values. The genes with p-value less than 0.01 are treated as an efficient gene and their expression value is to be downloaded. The gene that does not have gene name is required to filter out.
- **Relationship Identifying between Gene Pairs and GRN Generation** is focused on identifying relationship among each gene pairs and generating the gene regulatory networks (GRNs). First of all we use the singular value decomposition (SVD) for factorization of gene expression matrix. It brings related gene more close and unrelated genes more away. Thereafter, interactions between gene pains are generated by calculating Pearson's correlation coefficient. Since here we have interest to generate undirected weighted graph, therefore we takes absolute value of the correlation coefficients.

- **Clustering using MCL** is focused on clustering of weighted undirected graph. We clusters the graph for different inflation parameter r. We choose the r neither too small nor too large. If r should be too small then there should be fewer clusters of larger cluster sizes. If r will be too large then there should be more clusters of smaller size.
- **Key Gene Identification and Verification** focused to identify key genes and verify them from existing literature. For this purpose we use the clustered GRNs and identify the hub genes that have maximum degree. Thereafter, the genes that present as hub genes for different value of r are declared as key genes. It is required to verify these genes from existing literatures.

Further details about these functionalities along with the experimental results are presented in the following sections.

3. EXPERIMENTAL SETUP AND RESULTS

In this section, we present our experimental setup and results. We have used the GEO accession viewer of NCBI to load microarray gene expression profile data using GEO accession number GSE4988 [3] and GSE26126 [14]. There are total 20 samples out of which 12 samples are taken from colorectal cancer patients and 8 from healthy donors in GSE4988. There were total 15552 genes expressions, but a large number of the genes do not have gene names. In general cases the expression value of genes in cancer samples are higher than that of in normal samples. Figure 2 show the expression value of gene 'DDX46' in normal and cancer samples. From figure 2 it is clear that many of the normal sample profile are down regulated.

In GSE26126, there are total 193 samples out of which 181 samples was un-cultured and 12 taken from cultured. There were total 98 normal tissue samples, 95 tumor tissue samples, and total 27578 genes.

After loading this data from GEO, we analyse it using GEO2R tool. GEO2R web tool perform the a number of statistical testing using the GEO query and limma R package. It sort the gene in ascending order of p-values. There are options to select different columns and save the results. There are a number of algorithms and parameters for analysis. We use the default parameters and Benjamini & Hochberg algorithm for analysis of the gene expression data.

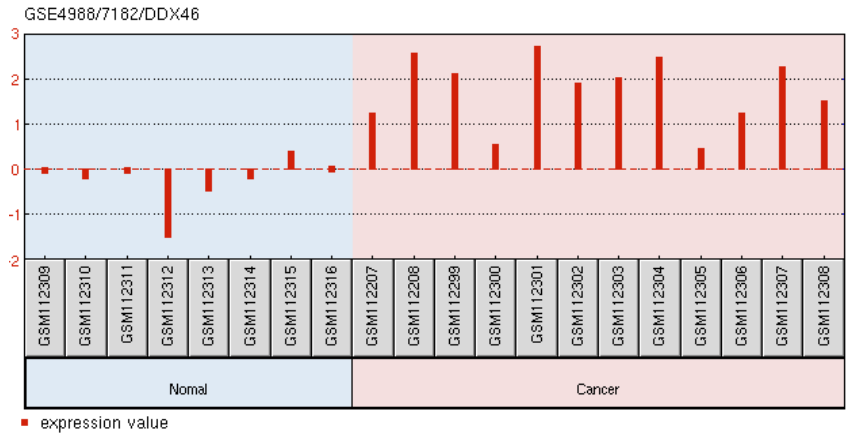

Figure 2: Expression Values of Gene 'DDX46'in Normal and Cancer Samples

There is a large number of genes, majority of them are irrelevant. After analysis, based on p-value we take the gene for further analysis. We select the gene from the analysed data whose p-value is less than 0.01. Thereafter, we eliminate the gene whose name is missing. Then we download the expression values of these genes for all samples.

From GSE4988 we gets 105 genes and their expression values, and from GSE26126 we gets 250 genes and their expression values.

From the extracted data we get a gene expression matrix A order m x n. Where a_{ij} element of this matrix represents the expression value of i^{th} gene in j^{th} sample, m is the number of genes and n is the number of samples. Now we apply the singular value decomposition (SVD) to factorize this matrix into – an orthogonal matrix U of order m × n, a diagonal matrix S of order n × n, and an orthogonal matrix V of order n × n. The rows of the matrix U represent the gene and columns represent the sample. Here each row vector of matrix U is corresponds to a gene. We apply the Pearson's correlation on rows of matrix U to get the relation between pair of genes. For these tasks we have written a java program. In gene regulatory network the nodes are corresponds to gene and edges are corresponding to their interaction. The GRN generated from GSE26126 is shown in figure 3, and clustered GRN is shown in figure 4 that is generated by applying MCL with r = 1.2.

From figure 4 it is clear that gene 'LTC4S' is hub gene. We cluster the GRNs of both data sets for different value of r and then identify the key genes that are hub genes in large number of clustered GRNs. The summary of these results are show in table 1.

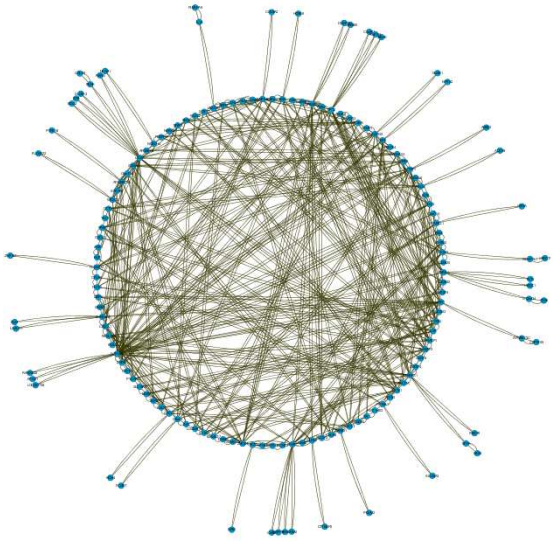

Figure 3: Gene Regulatory Network from GSE26126

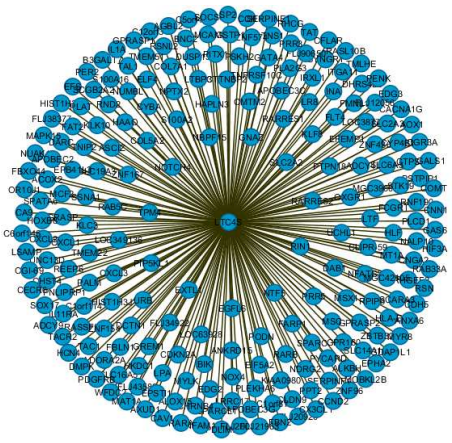

Figure 4: Clustered Gene Regulatory Network from GSE26126 by Applying MCL with R = 1.2

From the table it is clear that the gene 'LTC4S' and 'MYD88' are key genes in GSE26126 and GSE4988 respectively. In NCBI it is reported that MYD88 gene encodes a cytosolic adapter protein that play a central role in the innate and adaptive immune response [15]. In literature it is reported that LTC4S gene to be associated with the development of aspirin-induced asthma [16].

Table 1: Hub Genes from GSE4988 and GSE26126 Identified by MCL with Different r-Values

R	No. of Clusters of More than Two Nodes	No. of Nodes	No. of Edges	No. of Isolated Nodes	Hub Genes
GSE26126					
1.2	1	229	228	0	LTC4S
1.8	1	229	222	6	LTC4S
2.2	1	229	151	77	LTC4S
2.8	0	229	0	229	---
GSE4988					
1.2	1	104	103	0	MYD88
1.8	1	104	103	0	MYD88
4.8	2	104	96	5	MYD88, CBX1
5.2	2	104	77	22	MYD88, CBX1
5.8	3	104	39	56	MYD88, ACP2, HNRNPH3
6.8	1	104	14	81	MYD88

4. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented an application of graph clustering approach to identify key genes from microarray gene expression data. To this end, we have taken two gene expression data of GEO accession numbers GSE4988 and GSE26126. The genes of these data are analyzed using statistical techniques and sort on ascending order of their p-values. Genes having p-values less than 0.01 are considered and genes having no gene names were filtered out. As a result, total number of 105 and 250 genes are identified from GSE4988 and GSE26126 respectively, and their expression values for all the samples are downloaded. Thereafter, singular value decomposition for matrix factorization is used which brings related gene more closer and unrelated genes farther. The weighted undirected graphs for both data are generated by calculating the Pearson's correlation between gene pairs. Thereafter, MCL is applied on this weighted undirected graph to cluster it and the genes corresponding to hub node is identified as hub gene. The graph are clustered with different r values and finally the hub genes that are presents in almost all the clustered GRNs are identified as key genes. The 'MYD88' and 'LTC4S' genes are declared as key genes in GSE4988 and GSE26126 respectively as they are present in almost all the clustered gene regulatory networks, and on analysis it is found that these two genes play important role in in a number of diseases. In future, we have planned to extend our experiment on a larger number of data.

REFERENCES

1. Butte, A. J., & Kohane, I. S. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements, in Proceedings of *Pacific Symposium on Biocomputing 2000*, p. 418–429.
2. Caporaso, J. G., et al (2006). Concept recognition, information retrieval, and machine learning in genomics question-answering, in Proceedings of the *15th Text Retrieval Conference (TREC'2006)*, Gaithersburg, Maryland.

3. Collado, M., et al (2007). Genomic profiling of circulating plasma RNA for the analysis of cancer, *Clin Chem 2007 Oct;* vol. 53(10), pp. 1860-1862. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4988
4. Forsythe, G. E., et al (1997). Computer Methods for Mathematical Computations, *Prentice Hall Professional Technical Reference,* ISBN: 0131653326.
5. Jong, H. D. (2002). Modeling and simulation of genetic regulatory systems: a literature review, *Journal of computational biology*, vol. 9(1), pp. 67-103.
6. Karlebach, G., & Shamir, R. (2008). Modelling and analysis of gene regulatory networks, *Nature reviews. Molecular cell biology*, vol. 9(10), pp. 770-780.
7. Madhamshettiwar, P. B., et al (2012). Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets, *Genome Medicine*, vol. 4(41).
8. Maetschke, S. R., et al (2013). Supervised, semi-supervised and unsupervised inference of gene regulatory networks, *arXiv: 1301.1083v1*.
9. Mohapatra, S. K., & Krishnan, A. (2011). Microarray data analysis, *Methods in Molecular Biology*, Vol. 678, pp. 27-43.
10. Margolin, A. A., et al (2006). ARACNE: an Algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context, *BMC Bioinformatics*, vol. 7(Suppl. 1), pp. S7.
11. Wang X., & Gotoh, O. (2009). Microarray-based cancer prediction using soft computing approach, *Cancer informatics*, vol. 7, pp. 123-139.
12. Qiu, P., et al (2009). Fast Calculation of Pairwise Mutation for Gene Regulatory Network Reconstruction, Computer Methods and Programs in Biomedicine, vol 94(2), pp. 177-180.
13. Dongen, S. V. (2000). A cluster algorithm for graphs, University of Utrecht.
14. Kobayashi, Y., et al (2011). DNA methylation profiling reveals novel biomarkers and important roles for DNA methyltransferases in prostate cancer, *Genome Res*, vol 21(7), pp. 1017-1027. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE26126
15. Hawn, T.R., et al (2015). Myeloid Differentiation Primary Response Gene (88) — and Toll-Like Receptor 2—Deficient Mice Are Susceptible to Infection with Aerosolized Legionella pneumophila, *The Journal of Infectious Diseases*, Vol 193(12), pp. 1693-1702.
16. Kawagishi, Y., et al (2002). Leukotriene C4 synthase promoter polymorphism in Japanese patients with aspirin-induced asthma, *Journal of Allergy and Clinical Immunology*, Vol 109(6), pp. 936–942.